Data representation -

Graphs

One way to represent a graph is by adjacency matrix

each column represents a node and each row represents a node

cell i, j stores 1 if edge i-j exists in graph G, otherwise, it stores 0

	A	B	C
A	0	1	1
B	1	0	1
C	1	1	0

Another way to represent a graph is by adjacency list

A: (B, C)

B: (A, C)

C: (A, B)

stores lists of neighbors for each node

Images

matrix of pixels or list of pixels

Text

List of words, we want to keep track of these words and understand whether two documents are similar or dissimilar based on words they're using, vocabulary, are they talking about the same subject, concept (high level things you can't describe in a mathematical way)

Strings

interested in particular order, like in DNA

Time series

list of data at specific intervals of time, like in videos, temperature or simple values as afunction of time, like our heat and metal plate example from lecture 0

Types of Learning

Two types of learning

unsupervised learning
first portion of course

supervised learning
latter half of course

Supervised Learning

Tools that have some kind of prediction in mind
ex: finding correlation between cricket chirps and temperature

- predicting the number of cricket chirps given a temperature or predicting temperature given number of cricket chirps
- linear model telling us what is the expected number of cricket chirps based on the dots
- commonly referred to as regression

ex: age vs tumor size, want to find out whether tumor is malignant or benign

- for a new point, given an age and a tumor size, do i have a malignant or benign tumor?
- this type of learning is commonly called classification

Unsupervised Learning

less interested in specific model and predictions of the data, but more interested in inherent structure in the data
ex: clustering, where we would assign each point to a cluster
ex: given a collection of articles, do they cover the same topics?

- what are the topics covered by these articles?
- the concept of topics are vague and high level, but if we look at a set of articles in a specific way, the similarities between these topics shows up in a very clean structure that we can often assign to a high level concept or topic that we already know of, so we can see which are similar and talking about the same topic and so on.

Distance and Similarity

Data

n by m matrix, n data points and m features
every row is a specific data point that has m attributes
ex: name, age, balance (name doesn't really matter)

- what is the feature space for age and balance? what are all the values that age can have and what are all the values that balance can have?
- looks like a 2d plane
- we can plot our Jane and John vectors in this feature space

Distance

one of the key points when we do unsupervised learning, one of the fundamental things we want to look at is whether two points are similar or dissimilar, to compare two data points

dissimilarity function - a function that tells us whether two objects are dissimilar. large value means dissimilar and small value means similar.

distance function is a special type of dissimilarity function, so a dissimilarity function does not necessarily have all the constraints of a distance function

d is a distance function if and only if

- d(i, j) = 0 if and only if i = j
- d(i, j) = d(j, i)
- d(i, j) ≤ d(i,k) + d(k, j) — triangle inequality, shortest path from i to j

you don't need a distance function to compare data points, but we would prefer using a distance function because it's intuitive, easy to understand, easy to graph.

Minkowski Distance

For x, y points in n-dimensional real space
i.e x = $[x_1, \dots, x_n]$ and y = $[y_1, \dots, y_n]$
$p \geq 1$
$L_p(x, y) =$

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

When p = 2 → Euclidean Distance

When p = 1 → Manhattan Distinace

ex:
d = 2
p = 2 — distance between A (0,0) and B (1,1) is sqrt 2
p = 1 — distance between A and B is 2 (called Manhattan distance because everything is a grid, and you can only move along the lines to get from A to B)

d = 3
p = 1 — distance is 3
p = 2 — distance is 2 sqrt 2
p = 3 — distance is sqrt 3

Is L_p a distance function when $0 < p < 1$?
D(B,A) = D(A,C) = 1
$D(B,C) = 2^{1/p}$
But… if $p < 1$ then $1/p > 1$ which violates triangle inequality